

# テキストチャットを用いた雑談対話コーパスの構築と 対話破綻の分析

東中竜一郎<sup>†</sup>・船越孝太郎<sup>††</sup>・荒木 雅弘<sup>†††</sup>・  
塚原 裕史<sup>††††</sup>・小林 優佳<sup>†††††</sup>・水上 雅博<sup>††††††</sup>

対話システムが扱う対話は大きく課題指向対話と非課題指向対話（雑談対話）に分けられるが、近年 Web からの自動知識獲得が可能になったことなどから、雑談対話への関心が高まってきている。課題指向対話におけるエラーに関しては一定量の先行研究が存在するが、雑談対話に関するエラーの研究はまだ少ない。対話システムがエラーを起こせば対話の破綻が起り、ユーザが円滑に対話を継続することができなくなる。しかし複雑かつ多様な内部構造を持つ対話システムの内部で起きているエラーを直接分析することは容易ではない。そこで我々はまず、音声誤認識の影響を受けないテキストチャットにおける雑談対話の表層に注目し、破綻の類型化に取り組んだ。本論文では、雑談対話における破綻の類型化のために必要な人・機械間の雑談対話コーパスの構築について報告し、コーパスに含まれる破綻について分析・議論する。

キーワード：対話システム，非課題指向対話，雑談，破綻，エラー分析

## Text Chat Dialogue Corpus Construction and Analysis of Dialogue Breakdown

RYUICHIRO HIGASHINAKA<sup>†</sup>, KOTARO FUNAKOSHI<sup>††</sup>, MASAHIRO ARAKI<sup>†††</sup>,  
HIROSHI TSUKAHARA<sup>††††</sup>, YUKA KOBAYASHI<sup>†††††</sup> and MASAHIRO MIZUKAMI<sup>††††††</sup>

In general, there are two types of dialogue systems: the task-oriented dialogue system and the non-task-oriented or chat dialogue system. In recent years, chat dialogue systems have received much attention mainly because of the advances in automatic knowledge acquisition from the web. Nevertheless, few studies are dedicated to the error analysis of chat dialogue systems. This is in contrast with the many error-analysis-related studies on task-oriented dialogue systems. An error in a chat dialogue system can lead to the dialogue breakdown, where users are no longer willing to continue the conversation. Therefore, error analysis is crucial in such systems. However, it is difficult to analyze errors in chat dialogue systems because of the complex internal structures of the systems. In the present study, we analyze and categorize the errors

---

<sup>†</sup> NTT メディアインテリジェンス研究所, NTT Media Intelligence Laboratories

<sup>††</sup> (株) ホンダ・リサーチ・インスティテュート・ジャパン, Honda Research Institute Japan Co., Ltd.

<sup>†††</sup> 京都工芸繊維大学, Kyoto Institute of Technology

<sup>††††</sup> (株) デンソーアイティーラボラトリ, Denso IT Laboratory, Inc.

<sup>†††††</sup> (株) 東芝, Toshiba Corporation

<sup>††††††</sup> 奈良先端科学技術大学院大学, Nara Institute of Science and Technology

in a text chat dialogue system on the basis of the surface form of the conversations. We construct a chat dialogue corpus between a chat system and users and analyze the dialogue breakdowns included in the corpus.

**Key Words:** *Dialogue System, Non-Task-Oriented Dialogue, Chat, Breakdown, Error Analysis*

## 1 はじめに

近年 Twitter による人間同士の短文のやりとりを始めとしたインターネット上の大量の会話データから自動知識獲得 (稲葉, 神園, 高橋 2014) が可能になったことや, 高性能な音声認識機能が利用可能なスマートフォン端末を多くの利用者が所有するようになったことで, 雑談対話システムへの関心が, 研究者・開発者側からも利用者側からも高まっている。

対話システムが扱う対話は大きく課題指向対話と非課題指向対話に分けられるが, 雑談は非課題指向対話に分類される。課題指向対話との違いについていえば, 課題指向対話では対話によって達成する (比較的) 明確な達成目標がユーザ側にあり, 一般に食事・天気など特定の閉じたドメインの中で対話が完結するのに対し, 雑談では, 対話すること自体が目的となり, 明確な達成目標がないなかで多様な話題を扱う必要がある。また, 課題指向対話では基本的に対話時間 (目標達成までの時間) が短い方が望ましいのに対し, 雑談ではユーザが望む限り対話を長く楽しめることが望まれる。そのため, 適切な応答を返すという点において, 雑談対話システムは, 課題指向対話とは異なる側面で, 様々な技術的困難さを抱える。

これまで, 雑談対話システムの構築における最も大きな技術的障壁の 1 つは, 多様な話題に対応する知識 (応答パターン) を揃えるコストであった。上記のように, この問題はインターネットからの自動獲得によって解消されつつある。また, ユーザを楽しませる目的 (Wallace 2004; Banchs and Li 2012; Wilcock and Jokinen 2013) だけであれば, システムがおかしな発言をしてしまうことを逆手にとって, 適切な応答を返しつづける技術的な困難さを (ある程度) 回避してしまうことも可能である。

その一方で, 雑談対話には, ユーザを楽しませるという娯楽的な価値だけでなく, ユーザとシステムとの信頼関係の構築 (Bickmore and Cassell 2001) や, ユーザに関する情報 (ユーザの好みやユーザの知識の範囲) をシステムが取得することでユーザによりよいサービスを提供することを可能にする (Bang, Noh, Kim, and Lee 2015), 遠隔地にいる高齢ユーザの認知・健康状態を測定したり認知症の進行を予防する (小林, 山本, 大内, 長, 瀬戸口, 土井 2011), グループ内のコミュニケーションを活性化し人間関係を良好にする (Matsuyama, Akiba, Saito, and Kobayashi 2013), といった工学的・社会的価値が存在する。このため, 情報爆発, 少子高齢化, 生活様式の多様化と急激な変化による人間関係の複雑化といった諸問題を抱える現代社会において, 雑

談対話技術の更なる高精度化, すなわち適切な応答を返しつづける能力の向上が今まで以上に求められている。

雑談対話の高精度化のためには, 現状の技術の課題をエラー分析によって特定することが必要である。しかしながら, 課題指向対話, 特に音声対話システムにおける, 主に音声誤認識に起因するエラーに関しては一定量の先行研究が存在するが, テキストのレベルでの雑談対話に関するエラーの研究はまだ少なく, エラー分析の根本となる人・機械間の雑談対話データの蓄積もなければ, そのデータに含まれるエラーを分析するための方法論・分類体系も十分でない。

雑談対話システムがその内部でエラーを起こせば対話の破綻が起こり, ユーザが円滑に対話を継続することできなくなる。しかし, 対話システムは, 形態素解析, 構文解析, 意味解析, 談話解析, 表現生成など多くの自然言語処理技術の組み合わせによって実現され, かつシステム毎に採用している方式・構成も異なるため, システム内部のエラーを直接分析することは困難であるし, システム間で比較したり, 知見を共有することも容易ではない。そこで我々はまず雑談対話の表層に注目し, 破綻の類型化に取り組んだ。本論文では, 対話破綻研究を目的とした雑談対話コーパスの構築, すなわち人・機械間の雑談対話データの収集と対話破綻のアノテーションについて報告する。そして, 構築したコーパスを用いた分析によって得た破綻の分類体系の草案を示し, 草案に認められる課題について議論する。

以降, 2 節で対話データの収集について説明する。今回, 新たに対話データ収集用の雑談対話システムを 1 つ用意し, 1,146 対話の雑談対話データを収集した。3 節及び 4 節では, 上記の雑談対話データに対するアノテーションについて述べる。24 名のアノテータによる 100 対話への初期アノテーションについて 3 節で説明し, その結果を踏まえて, 残りの 1,046 対話について, 異なりで計 22 名, 各対話約 2 名のアノテータが行ったアノテーションについて 4 節で説明する。5 節では, 4 節で説明した 1,046 対話に対するアノテーション結果の分析に基づく, 雑談対話における破綻の類型について議論する。6 節で関連研究について述べ, 7 節でまとめ, 今後の課題と展開を述べる。

## 2 雑談対話データの収集

本研究は, Project Next NLP の対話タスク (関根 2015) の活動の一部として行われた。そのため, データ収集も対話タスクの参加者を中心に行った。本タスクに参加したのは, 表 1 に示す大学・企業を含む 15 の拠点からの総勢 32 名である。これは, 対話システムに関する国内のプロジェクトとして最大級の規模である。

雑談対話の収集は, 本研究のために新たに設けた専用の Web サイト<sup>1</sup>で行った。この Web サ

<sup>1</sup> <http://beta.cm.info.hiroshima-cu.ac.jp/~inaba/projectnext/>

イトでは、NTT ドコモが一般公開している雑談対話 API(大西, 吉村 2014)<sup>2</sup> を用いた雑談対話システムが稼動しており、Web ブラウザでアクセスすることで、テキストでの雑談を行える。このサイトでは、ユーザが 10 発話を入力すると対話が終了し、対話ログが出力されるようになっている。サイト側ではユーザ管理を行っておらず、ユーザが自己の対話を纏めて後日提出することによって、ユーザと対話ログの対応が取れるようになっている。図 1 に雑談対話収集サイトのスクリーンショットを示す。

各拠点の参加者および参加者の知人にこの雑談対話システムと対話をしてもらい、全部で 1,146 対話を収集した。収集は 2014 年 8 月 2 日から 31 日の間に行った。対話をする際には、図 2 に示す「対話ガイドライン (10 箇条)」に沿うように教示をした。これは、現状の雑談対話システム技術が成熟していないこともあって、破綻だらけの対話ばかり収集されてしまわないためである。

ユーザ毎の対話数は 1 から 40 までばらつきがあるが、全体のユーザ数は 100 名を超えている。表 2 は収集された雑談対話データの統計情報である。収集した雑談対話の一例を以下に示す。

表 1 対話タスクの参加状況

参加拠点	秋田県立大学, デンソーアイティーラボラトリ, 広島市立大学, ホンダ・リサーチ・インスティテュート・ジャパン, 京都工芸繊維大学, 京都大学, 名古屋工業大学, 奈良先端科学技術大学院大学, NTT, 大阪大学, 首都大学東京, 東芝, 豊橋技術科学大学, Yahoo!, 山形大学
参加者数	32

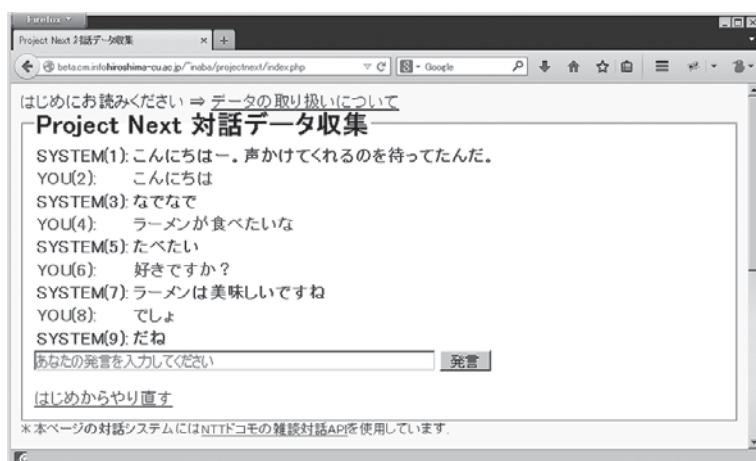


図 1 雑談対話収集サイトの画面

<sup>2</sup> [https://www.nttdocomo.co.jp/service/developer/smart\\_phone/analysis/chat/](https://www.nttdocomo.co.jp/service/developer/smart_phone/analysis/chat/)

- (1) たまたま待合室や飛行機などで隣り合った見知らぬ人と話すイメージで対話しましょう。特定の個人を想定して対話を行わないように注意してください。
- (2) システムの発話はなるべく好意的に解釈し、対話が続けるように努力しましょう。システムの力量を試すような発話は控えてください。また、長文の入力や、人間でも答えにくいような難しい問いかけは避けてください。
- (3) 対話毎に新しい気持ちで話しましょう。残念ながら、システムは前のあなたとの対話を覚えていません。新しい対話セッションでは、前回の対話のことは忘れて対話してください。
- (4) 自分から話題を開始しましょう。システムは、話題となる単語（主に名詞）があると、その単語に基づいて対話を行うことができますが、そのような単語が見つからないとうまく対話ができないことがあります。システムから話題を振られるのを待つのではなく、自分から話題を振りましょう。なお、マニアックな話題にはうまく対応できないことがあります。
- (5) 自分から話題を変えましょう。システムは、自分から話題を変えることをあまりしません。対話が進展しなかったり、システムがよく分からない応答を繰り返したりするときは、自分から新しい話題に変えてみましょう。
- (6) 何度も同じことを繰り返さないようにしましょう。システムの理解力が対話中に突然改善することはありません。ある発話について満足のいく回答が得られなくても、システムに対して、同じ質問や発話を繰り返したりしないようにしましょう。
- (7) 誹謗中傷は控えましょう。ここで収集しているデータは公開されます。データを見た人が不快になるような発言は避けて下さい。誹謗中傷とは、システムに対するもの、特定の個人に対するものの両方を含みます。
- (8) 個人情報を入力しないようにしましょう。公開されるデータであることを念頭において、名前、住所、電話番号などの個人情報は発話に含めないで下さい。第三者の個人情報の入力は絶対にやめてください。
- (9) 日本語で入力しましょう。システムは日本語しか理解できません。また、大阪弁や博多弁などの方言にも対応していませんので、標準語で入力してください。顔文字・フェイスマークなども理解できませんので使用しないでください。
- (10) 上記を守っていれば、あとは自由です。対話を楽しみましょう。

図 2 対話ガイドライン（10 箇条）

表 2 収集された雑談対話データの統計情報

対話数	1,146	
話者数	116	
	ユーザ	システム
発話数	11,460	12,606
異なり発話数	10,452	7,777
単語数（形態素数）	86,368	76,235
異なり単語数	6,262	5,076

形態素解析には MeCab の version 0.996（辞書は ipadic-2.7.0-20070801）を使用した。システム発話数は 12,606 となっているが、これは最初のプロンプト発話の 1,146 とユーザへの返答の 11,460 を合わせた値である。

S はシステム発話、U はユーザ発話を表す。

S: こんにちは！声かけてくれるのを待ってたんだ。

U: 今日は最高気温 36 度だって。暑いねえ

S: 最高気温は 17 度が予想されます??

U: いやいや猛暑ですよ

S: 猛暑に備えてますか?

U: うん、日傘さしたり、水飲んだり、クーラー入れたりしてる



- S: クーラーは快適ですね  
U: うん、会社だと朝はつけられないから死にそう  
S: 会社は冷房が効きますよねえ  
U: 昼間は効くね。ここはパソコンがいっぱいあるからないと大変  
S: 動きが悪いです  
U: そうなの？ こっちは快適だよ  
S: パソコンは使い勝手が悪いです

### 3 初期アノテーション

収集した雑談対話データ全体に対して対話破綻をアノテーションする前に、その一部について初期アノテーションを行った。具体的には、1,146 対話からランダムに 100 対話を抽出し、タスク参加者の中で全 100 対話のアノテーションを行える 24 人によってアノテーションを行った。作業期間は 2014 年 10 月 7 日から 17 日の間である。このアノテーションの目的は、残りの 1,046 対話に対して、1 対話あたり何人のアノテータを割り当てるのが妥当かを検討することである。ここで作成したデータセットのことを以後 **init100** と呼ぶ。

アノテーションについては、どのようなエラーがあるのかを網羅的に分析したいという目的に鑑み、トップダウンな破綻の分類は示さず、直感に従って○・△・×の3分類でアノテーションするように指示した。それぞれの意味は以下の通りである。

- 破綻ではない：当該システム発話のあと対話を問題無く継続できる。
- △ 破綻と言い切れないが、違和感を感じる発話：当該システム発話のあと対話をスムーズに継続することが困難。
- × 明らかにおかしいと思う発話（破綻）：当該システム発話のあと対話を継続することが困難。

多人数でアノテーションする場合には、○×の判断の分布によりそれらの中間状態を表現できるため、必ずしも△のような中間レベルを表すカテゴリを用意する必要はないが、アノテータが○か×かを迷うケースで判断に時間がかからないようにする目的で△を導入した。

アノテーションには、図 3 に示す専用のツールを使用した。ツールでは、非文のチェックの他に、各発話に対してコメントを記入できるようになっている。また、先行する文脈のみに基づいて対話破綻のアノテーションが出来るように、1 発話アノテーションする毎に、次のユーザ発話とシステム発話が表示されるようになっている。なお、破綻（△あるいは×）とタグをつけた後の発話をどうアノテーションするかについては、対話の先頭から、破綻とタグ付けされた発話を含むこれまでの文脈を「ありき（与えられたもの）」として、アノテーションするように教示した。すなわちアノテータは、破綻があったところで対話がリセットされたとはせず、

破綻も含めて先行文脈として作業を行った。

非文の定義は、「文法エラーなどにより日本語としての意味をなさない文」とし、会話体で許可される程度の「助詞落ち」や「ら抜き」は非文に該当しないとした。また、全く意味が通らない発話であれば当然×を付けることになるが、非文であっても発話意図が汲み取れるのであれば、○や△を付けてもよいとした。

### 3.1 非文の割合

使用した対話システム (大西, 吉村 2014) の応答生成は、人がすべて確認したテンプレートによるものではないので、非文の発生を完全に無くすることはできない。そこで、アノテーション時の非文のチェックの結果に基づき、文法レベルでの対話コーパスの品質を確認しておく。

最初のプロンプトを除くシステム発話全 1,000 発話において非文のチェックが付けられた発話の分布を表 3 に示す。表 3 の 1 行目は、ある発話に対して非文と判断したアノテータの数を



図 3 雑談データ用破綻アノテーションツール

表 3 init100 における非文の分布

人数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	合計
発話数	66	13	7	9	6	4	3	3	3	0	4	2	3	1	1	0	1	0	1	127

表す。2行目は各人数のアノテータに非文と判断された発話の数を表す。1人でも非文と付けた発話は1,000発話中127あったが、過半数（13人以上）が非文と付けたものはわずか7発話しかなかった。実際のデータを見ると、非文と判定したのが数名である発話はどれもアノテーション指示者からみて非文と判断するようなものではなかった。（「ルールは多いです」「価値観は欲しいです」「話し相手に飢えます」など、不自然ではあるが『日本語としての意味をなさない文』とまでいえない発話であった）。仮に過半数以上が非文としたものを真の非文とし、それ以外をアノテーションの誤りとすれば、init100での非文の発生率は1%未満である。

今回の初期アノテーションでは、24人全員が100対話を同じ順序でアノテーションしている。その中で最も非文であると判定したアノテータが多いシステム発話は「熱中症に気をつけか??」というものであった。この発話は100対話中で4回発生しており、4発話に対して非文とチェックした人数は、出現順で19人、17人、13人、9人であった。つまり過半数が破綻と付与した7発話のうち3発話は同一の発話であった。

同一内容の発話に対して「非文」とアノテーションした人数が大きくばらついているのは、既に非文と付けた発話に対する非文のチェックをアノテータが省略したことが原因と思われる。非文のチェックは、任意とも指示していないが、厳守するようにも指示しなかった。また、非文のチェックボックスは任意入力フィールドであったコメント欄の直前に置かれていた（図3参照）。このため、非文のチェックがアノテーションの主たる目的ではなく補助的な作業であったことから、後の方になるほどチェックを省略されてしまった可能性が高い。その事を考慮して、仮に四半（7人）以上が非文と判定したものを「真の非文」と考えても、非文の発生率はおよそ2%である。このことから、今回のデータ中のシステム発話の品質は、個々の発話の日本語文法のレベルでは、当面の研究に必要なレベルが担保されていると考える。

### 3.2 アノテータ間の一致度の分析

init100に対して、24人のアノテータが付与したラベル○、△、×の割合を表4に示す。24人のアノテータ間の一致の程度を測るために Fleiss の  $\kappa$  を算出すると、0.276であった。（Landis and Koch 1977）も参考にすると、この値の解釈は「ランダムではないが、よく一致しているともいえない」とするのが妥当である。△を×に含めて、2値のアノテーションとして計算すると、0.396とやや一致の具合が高まる。△を○に含めると  $\kappa$  は0.277にしか改善されないため、△は×により近いことが分かる。

表4 init100中の○△×の発生割合（発生数）

○	△	×
59.2% (14,212)	22.2% (5,322)	18.6% (4,466)



24 人のアノテータを Cohen の  $\kappa$  値をもとに Ward 法で階層クラスタリングを行うと, 図 4 のようになった. 距離の定義やクラスタリングの手法を変えると, 2 つのクラスタの中でのまとまり方は細かく変わるものの, 大きな 2 つのクラスタ間での移動はほとんど見られなかった. 図 5 に示す 24 人のアノテータの分布を見ると, ○をつける傾向の大小で, 前述の 2 クラスタが分かれていることが見て取れる. 2 つのクラスタの中での Fleiss の  $\kappa$  を求めると, それぞれ 0.414 (11 人) と 0.474 (13 人) であり, これらの値は「適度に一致している」と解釈できる. 前者, 図 4 の左側のクラスタ, を C1 と呼ぶ. このクラスタは○を多く付けるアノテータのクラスタである. 後者, 同右側のクラスタ, を C2 と呼ぶ. このクラスタは○を少なく付ける破綻に厳しいアノテータのクラスタである.

表 5 に, 24 人のアノテータの属性 (性別, 年齢層, 職業, 関係性) の分布を示す. 職業の「学生」は大学生および大学院生, 教員は大学教員を指す. 関係性の「当事者」は, 対話タスクに参加している研究者 (会社員, 教員, 学生) のことで, 関係者は, 対話タスクには直接参加し

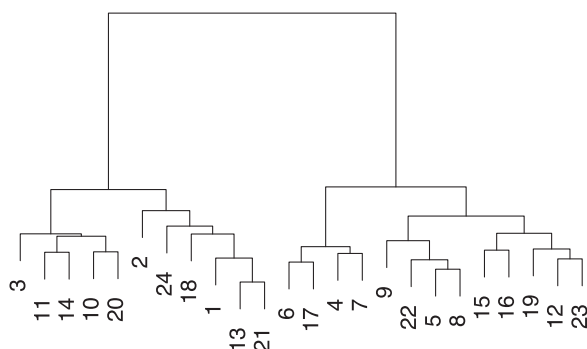


図 4 アノテータのクラスタリング結果 (番号はアノテータ ID)

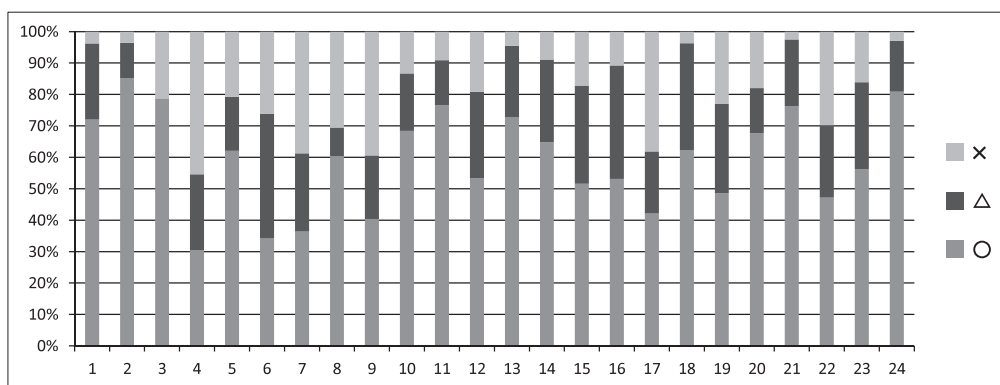


図 5 アノテータ毎の○△×を付与した割合 (横軸はアノテータ ID)

ていないが、前述の当事者と同じグループで対話システムに普段から関わりのある仕事をしていることを意味する。無関係は、当事者と知己であるが、対話システムの研究開発とは普段関わりがないことを指す。性別・年齢層には、C1 と C2 の間に目立った違いは見取れない。職業・関係性をみると、教員・当事者が C1 側にやや多い印象を受けるが、Fisher の正確確率検定では C1, C2 間に統計的に優位な差はない（いずれも  $p > .2$ ）。従って、表 5 に示した属性だけでは、新規のアノテータがどちらのグループに属するかを予測することは難しく、実際にアノテーションを行ってもらって傾向を把握するしかない。

24 人のアノテータからランダムに  $N$  人を選び出したとき、ラベルの分布がどれだけ全体の分布から離れているのかを表したグラフを図 6 に示す。横軸は  $N$  の数で、縦軸は Kullback-Leibler divergence の対称平均の値である。黒丸が 1,000 回サンプリングした際の平均値を示す。下向き三角は 1,000 回中の最大値、上向き三角は 1,000 回中の最小値を表す。アノテータが 1 人から 2 人になる段階で、平均値からの乖離は半分近く縮まり、あとは、なだらかに 24 人の分布に近寄っていくことが分かる。

図 7 に、○△×の各ラベルを付与された数毎の発話数のグラフを示す。左側のグラフは集計結果をそのまま示したもので、横軸が、ある発話について付与された特定のラベルの数（0 から

表 5 アノテータの属性分布

クラス	人数	性別		年齢層				職業				関係性		
		男	女	20 代	30 代	40 代	50 代	学生	教員	会社員	その他	当事者	関係者	無関係
C1 (弛)	11	10	1	5	4	1	1	4	5	2	0	9	2	0
C2 (厳)	13	10	3	7	5	1	0	5	2	5	1	7	5	1

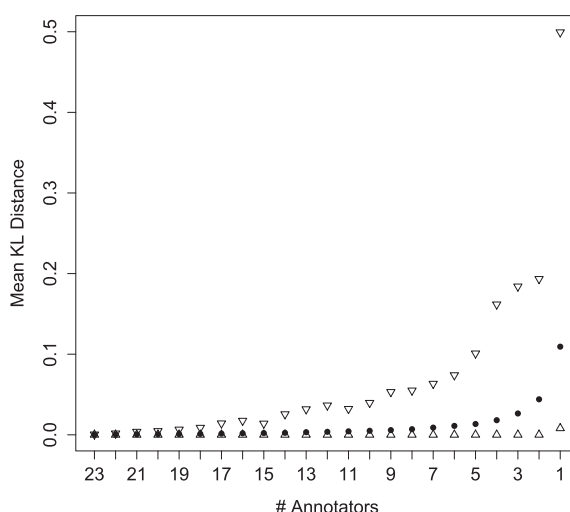


図 6 24 人のラベル分布とランダムサンプリングした  $N$  人のラベル分布の異なり

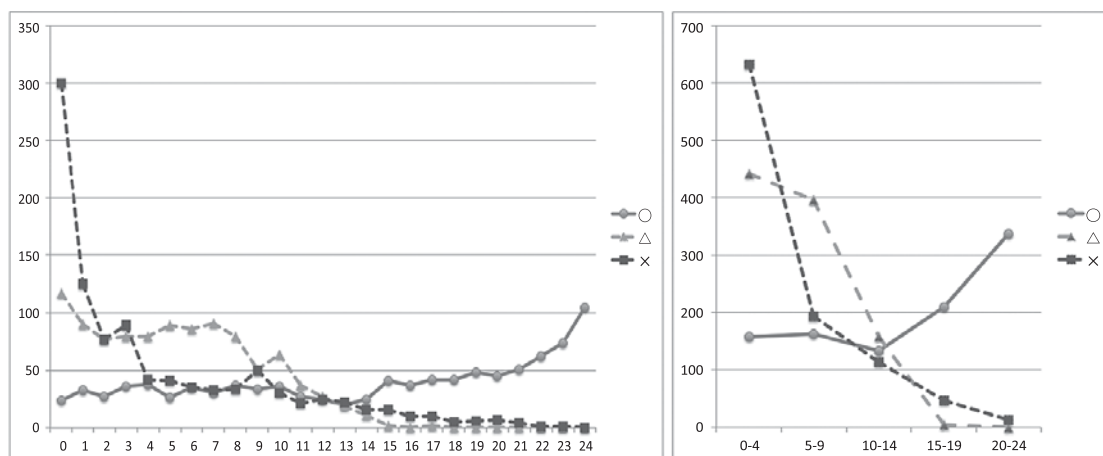


図 7 各ラベルを付与された数毎の発話数 (横軸：付与されたラベルの数, 縦軸：発話の度数)  
左図：元データ, 右図：5 区間に集約した結果

24) を表し, 縦軸が, そのような発話の度数を表す. 右側のグラフは, 解り易さのため, 左側のグラフを加工したものである. 0 から 24 までのラベルの数を 5 区間に区切り, 各区間の積算値をプロットしてある. このグラフから, 破綻ラベル×が過半数以上に付与されてる発話の数は少なく, 人により揺れが大きいことが読み取れる. また過半数以上が△を付与している発話はほとんどないことも分かる. ○については一致し易いが, △×について一致しにくいということを表している. △と×を併せて計数した場合のグラフは, 図中の○のグラフを左右反転させたものに等しくなる.

○の数と, △と×を併せた数とが 12 対 12 で半分に分かれた発話の分析も行った. すると, 表面的には間違ってるとは言えないけれども, 違和感がありそうな応答が多かった. 「こういうことは普通ここでは言わない」という印象は持つものの, なぜそれを言わないかの説明が難しいものが多い. また, 発言の意図が読みにくそうな発話, ユーザの話を聞いていないと感じられる発話, 社会的に問題がある発話, ふざけているようにも見える発話なども判断が割れていた. これらの判定を分析することは, 対話における認知の研究として学術的に面白い課題となり得る. 以下は, 評価が半分に割れた発話 (下線部) を含む対話の例である.

- U: 他に好きな楽器はありますか?  
 S: サックスがいいですね  
 U: サックスは木管楽器なんですよ  
 S: 相性がいいですね  
 U: 台風は大丈夫でしたか  
 S: 台風は必要です  
 U: 雨は必要ですよ

S: 朝から雨が降るのです

これは発話の意図が読みにくいいため、「それがどうしたのか」という印象を受ける発話である。

U: 夏って感じがします

S: 海は騒がしいですねえ

U: 海に行くと盛り上がりますもんね

S: 海は静かなんですねー

U: 騒がしいのか静かなのかどちらですか

S: 海はサイコーですねえ

U: サイコーですねえ

これは人によっては、質問に答えずにごまかしているように感じられるだろう発話である。

### 3.3 コメントに基づく破綻の分類

init100 中の一部の対話について、アノテータによって対話破綻箇所が付与されたコメントを人手で分類したところ、大きく「同じ内容の繰り返し」「矛盾した内容」「発話として唐突」「ユーザ発言内容の無視」「質問に答えていない」に分類できることが分かった。

**同じ内容の繰り返し** 表層としては少し異なっていたとしても内容として同じものを繰り返す場合、破綻とアノテーションされることが多かった。たとえば、「美味しいですね」「いいですね」などと同じような発話を繰り返す場合である。

**矛盾した内容** システム発話間で矛盾が見られる場合は破綻とされることが多かった。たとえば、「イチゴが好き」という発言の直後に「リンゴが好き」と発言するなど、一貫性を欠く発話は問題視された。

**発話として唐突** 「おはようございます」に対して「明けましておめでとうございます」のように、文脈とは関係のない発言を突然行うことがあり、このような発話は破綻とされていた。

**ユーザ発言内容の無視** 対話はお互いが協調して進めていくものであるもので、ユーザ発話を全く受けずにシステムが発話を行った場合には対話の破綻とみなされることが多かった。たとえば、旅行の話をしていて「車で行きましょう」とユーザが話しかけたのに「車はカッコいいですね」と車そのものについて言及したりする場合である。

**質問に答えていない** ユーザ発言内容の無視に近いが、特に質問に答えていないものが破綻とされていた。たとえば、「チワワは欲しいですね」とシステムが話し、それに応じてユーザが「飼う予定はあるの?」と質問したが、システムは「チワワはいいらしいですよ」と答えたような場合である。

上記以外にも口調の唐突な変化などが、問題のある現象として観察された。さらに詳しい分類については5節で述べる。

## 4 残りの対話へのアノテーション

init100 に対するアノテーション結果について, タスク参加者で議論を行った結果, 残りの 1,046 対話 (以後, **rest1046** と呼ぶ) のアノテーションについては, 1 対話につき 2 人で実施するという結論に至った. 2 名とした理由は以下の通りである.

- 人的・経済的コストの面から, アノテーションにかかる作業量は最小限が望ましい.
- アノテーションのコストを最小化できるのは 1 名でアノテーションを行う場合であるが, この場合, アノテータ間の揺れのために, 破綻とされるべき発話が見逃されてしまう可能性がある. よって, 複数名が望ましい.
- 前述の分析でアノテータは大きく 2 つのクラスタに分かれることが分かっている. これらの 2 つのクラスタから 1 名ずつ割り当てすることで, 見逃しを最も効率的に減らせる可能性がある.

実際に, init100 にアノテーションをした 24 人からランダムに  $N$  人をランダムに選んだ場合と, C1 と C2 の両クラスタから  $N/2$  人ずつ選び出した場合とで, 図 6 と同じ方法でラベル分布の距離を比較すると, 図 8 に示す結果になる. C1, C2 のクラスタから 1 人ずつ, 計 2 名選んだ場合の結果は, 全体からランダムに 3 人選んだ場合と 4 人選んだ場合の中間程度になっており, より少ない人数で全体での分布に近い結果を得られることが分かる.

1,046 対話をランダムに 11 個のサブセット (a-k) に分割した. a-j の 10 個のサブセットはそれぞれ 100 対話を含み, 最後のサブセット k だけが 46 対話を含む.

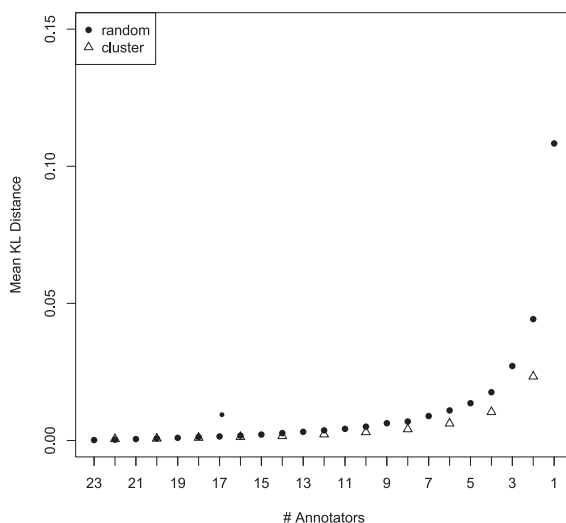


図 8 24 人のラベル分布とランダムサンプリングした  $N$  人のラベル分布の異なりの比較: 完全にランダムな場合 (random) と, クラスタ C1・C2 を考慮した場合 (cluster)



アノテーションには、22名のアノテータの協力が得られることになった。22名のうち19名が、init100に対するアノテーションに参加していたアノテータである。まずこの19名について、図4のクラスタに基づき、2つの大クラスタC1およびC2からなるべく1名ずつのアノテータが割り当てられるように、サブセットkを除く10サブセットに割り当てた。その後残りの3名を同10サブセットに割り当てた。1名当りの分担量を2サブセットと固定して22名を10サブセットに割り当てたので、i, jの2つのサブセットだけ3名のアノテータを割り当てた。サブセットkについては、余力のある2名に割り当てた。

アノテータが各対話にアノテーションを行う方法は、init100の場合(3節)と同じである。アノテーションの結果の分布を表6に示す。init100よりも、△の割合が増えているが、△と×を併せて見た場合には、init100のときとほぼ同じ分布と考えられる。また、各サブセット毎のFleissの $\kappa$ 値を表7に示す。2名のアノテータが同じ判断傾向を持つかどうかによって、サブセット間で $\kappa$ 値にばらつきが生じているが、全体平均としてはinit100とほぼ同じ値になっている。

rest1046全体について、2名のアノテータが付けたラベルの組み合わせ毎の頻度と割合を図9に示す(計算にあたりサブセットi, jの3人目のアノテーションは利用していない)。先に述べたように、アノテータは○を多く付ける傾向のクラスタC1と、そうでないクラスタC2とに大きく分かれており、各サブセットに割り当てるアノテータは、なるべく2つのクラスタから1名ずつ選ぶようにした。図9では、整合した判定である(×, ×)の組よりも、矛盾した判定である(○, ×)の方が数が多くなってしまっているが、これは上記の割当の結果を反映しているもので想定内の結果であると同時に、破綻の捉え方が人によって異なることを改めて示している。

rest1046のアノテーションに際しては、担当する対話の最初の5対話と最後の5対話、計10対話だけ、△, ×をつけた箇所には、必ずその判断理由をコメントとして書くことを求めた。こ

表6 rest1046中の○△×の発生割合(発生数)

○	△	×
58.30% (13,363)	25.33% (5,805)	16.37% (3,752)

表7 サブセットa-k毎のFleissの $\kappa$ 値(i, jのみ3名でのアノテーション、その他は2名ずつ)

サブセット 対話数	a 100	b 100	c 100	d 100	e 100	f 100	g 100	h 100	i 100	j 100	k 46	計 1,046
○の数	1,271	1,159	1,222	1,174	1,186	693	1,150	975	2,162	1,781	590	13,363
△の数	550	522	474	258	400	732	543	633	567	863	263	5,805
×の数	179	319	304	568	414	575	307	392	271	356	67	3,752
$\kappa$	0.31	0.38	0.19	0.30	0.37	0.36	0.23	0.14	0.24	0.29	0.27	0.28*

(\*マクロ平均)

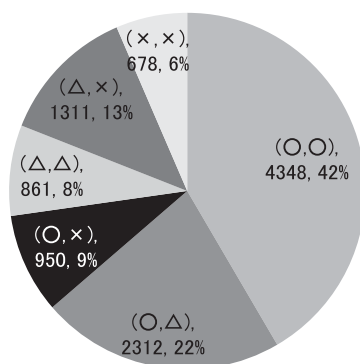


図 9 2 名のアノテータによるラベルの組み合わせの頻度と割合

れにより, 総数で 3,748 個, 異なりで 2,468 個のコメントを得た. アノテーション作業は 2014 年 12 月 2 日から 20 日の間に行った.

## 5 対話破綻の類型化

本節では, 収集したデータを基に策定を進めている対話破綻の分類体系の, 現時点での案と課題について議論する. 3 節では init100 に対して付与された △, × の破綻アノテーションに付随するコメントを大まかに分類した結果を示したが, ここではそれを土台としつつ, rest1046 に対して付与されたコメントを分析し, 雑談対話における対話破綻の類型化を行った結果を示す.

対話が, ある発話によって破綻するとき, 原因はその発話だけにあるとは限らない. もちろん, その発話が文法におかしなものであったり, 意味がわからなかったりする場合もある. しかし, その発話が文として正しいものであったとしても, 「相手の発話に対して, このように応答するのはおかしい」場合や, 「前に言ったことと矛盾している」という場合においても, 対話の継続が困難となる. このように, 対話の破綻を分析するに当たっては, 当該発話そのものに原因があるのか, または広い意味での文脈 (直前の発話, 対話履歴, 状況なども含む) に原因があるのかを特定する必要がある.

また, 破綻が生じた原因が存在する範囲が同じであっても, その内容は様々である. 必要な情報の欠落や曖昧性のために意味が特定出来ない場合や, 意味が特定できても文脈と矛盾する場合, 矛盾はしなくても冗長な場合などがある.

そこでまず, 破綻の根拠となっている情報に基づき大分類を決定し, その後, 破綻の種類を表す小分類を決定した. 大分類は, 破綻を認定する際にどの範囲に関連した破綻であるかという基準で, 以下の 4 つに決定した (図 10 参照).

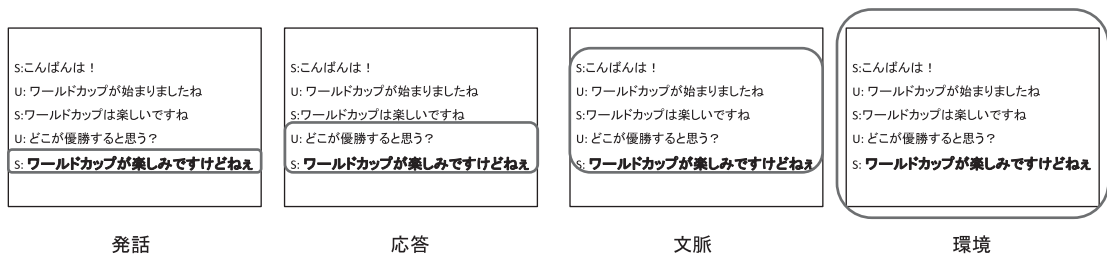


図 10 大分類を決める基準（範囲の違いを模式化した図であり，図中の発話は必ずしも各ケースに実際に該当する発話ではない）．太字は破綻と認定された発話．

- 発話

当該システム発話のみから破綻が認定できるケース．典型的には非文が該当する．「意味不明」というコメントの場合でも，この発話単独で意味がわからないのではなく，前の発話や文脈との関係で意味が取れない，というケースがあるので注意した．

- 応答

直前のユーザ発話と当該システム発話から破綻が認定できるケース．典型的には，発話対制約違反や，前発話の話題を無視した応答などが該当する．あくまでもそれまでの対話の流れは無視して，1つ前の発話との関係だけで判断した．

- 文脈

対話開始時点から当該システム発話までの情報から破綻が認定できるケース．典型的には，対話の流れから判断できる不適切な発話・矛盾する情報の提供・不要な繰り返しなどが該当する．

- 環境

破綻原因が，「環境」すなわち「外部要因」にあり，上記の3分類には当てはまらないケース．典型的には，一般常識に反するシステム発話が発話される．

## 5.1 対話破綻の分類体系案

表 8 に示す対話破綻の分類体系案を考案した．「発話」・「環境」の大分類については，検討の段階で多数を占めた「誤り」と分類される発話に対して，より分解能が高まるようにそれぞれ小分類を設定した．

一方，「応答」・「文脈」の大分類においては，(Bernsen, Dybkjær, and Dybkjær 1996; Dybkjær, Bernsen, and Dybkjær 1996) に倣い，対話における協調の原則である Grice の公準 (Grice 1975) に基づき小分類を設定した．Grice の公準は，量・質・関係・様態の各公準からなるもので，対話において参加者が遵守するように期待されている原則である．つまり，ユーザの直前の発話あるいはこれまでの対話履歴を受けてなされるシステム発話が守ると期待されている原則であ

表 8 分類体系草案

大分類	小分類	類型名	説明
発話	構文制約違反	構文的誤り	日本語の文として正しくない
	意味制約違反	意味的誤り	文としての意味が通らない
	不適切発話	解釈不能	著しく応答の体をなしていない
応答	量の公準違反	情報過不足	応答として内容の過不足がある
	質の公準違反	不理解	直前のユーザ発話を理解していない
	関係の公準違反	無関係	直前のユーザ発話の話題や発話意図と無関係
	様態の公準違反	意図不明	発話の意図が汲み取れない
	誤解	誤解	直前のユーザ発話の内容を誤って理解している
文脈	量の公準違反	不要情報	冗長な発話
	質の公準違反	矛盾	発話内容や態度が急転換する発話
	関係の公準違反	無関係話題	話題が文脈から逸脱
	様態の公準違反	関連性不明	文脈のどの部分と関連しているのか不明
	話題展開への不追随	不追随	話題展開後も前の話題を続けている
環境	無根拠	共通基盤欠如	根拠の無い主張
	矛盾	一般常識欠如	常識に反する主張
	非常識	社会性欠如	社会規範から外れ, 相手を不快にする発話

るので, 一般的にはこの原則が守られていないと, ユーザはシステムの発話意図を推測することができず, 対話が破綻すると考えられる. (Bernsen et al. 1996; Dybkjær et al. 1996) は, 課題指向対話のエラー分析に Grice の公準を用いて, 一定の成果を得ている. アノテータのコメントに「答えていない」「無視しすぎ」「唐突すぎる」といった「違反」を示唆するものが多かったことも, 対話の「規範」である Grice の公準を用いた理由の 1 つである. 破綻を分類することの一義的な目的は, ユーザが破綻であると考えた箇所で, システム内部のどこに問題 (エラー) があったのかを探ることであり, それを知ることによって, システムの改善が可能になる. しかしながら, 一般にシステム内部でおきたエラーを対話の表面から直接特定することは難しい. そこで, システムが何をしてしまったのか (どういう問題行動をしたか) を Grice の公準に基づいて分類することを基本とした. ただし, 問題行動の原因が比較的解り易いエラーについては, 「応答」大分類中の「誤解」および「文脈」大分類中の「不追随」として追加した.

以下で, 大分類毎に, 小分類について事例を基に説明する.

### 5.1.1 「発話」の小分類

基本的に, システム発話の生成過程に問題がある場合が該当する. 「非文」の概念に近い.

- 構文制約違反 (構文制約違反)

必須格の欠如や助詞の誤りなどが原因で, 発話内容が理解できないもの.

例) 「S: 熱中症に気をつか??」

- 意味制約違反（意味制約違反）  
意味素性の異なる語の組合せなどが原因で、発話内容が理解できないもの。  
例) 「S: 便乗に重ねますよねえ」
- 不適切発話（不適切発話）  
発話として機能を果たさないもの。  
例) 「S: ちゅっちゅ」

### 5.1.2 「応答」の小分類

前発話との関係において協調の原則が守られていない破綻を、Griceの公準に基づいて分類する。誤解は内部的なエラーであり、対話行動自体の問題ではないが、比較的解り易いので小分類として加えた。

- 量の公準違反（情報過不足）  
前発話に対して、応答に含むべき内容が欠落している、あるいは、不要な内容が含まれており、これが会話の進行を阻害する場合。  
例) 「U: ラーメンが好きです」→「S: 豚骨で有名ですね」
- 質の公準違反（不理解）  
前発話と矛盾する内容の発話。前発話の話題に追従しながらも、前発話で示された命題内容と矛盾する命題内容を含む場合。なお、意味理解の失敗などによって生じる「ずれ」を含む発話は、関係の公準違反とする。  
例) 「U: 最近、山登りを始めました」→「S: 山登りが出来ないのですか？」
- 関係の公準違反（無関係）  
発話対として不自然な発話で、これによって会話の継続ができなくなっている状況。あるいは、前発話の意味内容と関係しない内容の発話を行った場合、相手の発話を無視している状況を含む。  
例) 「U: それでは、趣味はなんですか？」→「S: 最後に旅行されたのはいつですか？」
- 様態の公準違反（意図不明）  
発話意図の曖昧性のため、会話の継続ができなくなっている状況。  
例) 「U: 退屈は嫌いですか？」→「S: 退屈は心ですね」
- 誤解（誤解）  
内容の解釈エラーによって、不適当な発話になっている状況。多義語の意味の取り違いなどが該当する。  
例) 「U: スイカを食べたらお腹を壊します」→「S: スイカで買い物できてるのですか？」



### 5.1.3 「文脈」の小分類

基本的には、「応答」の小分類と同様の考え方で、協調の原則に対する違反で分類する。

- 量の公準違反（不要情報）

与える情報が何もない無情報発話、あるいは、これまでの文脈に出現した発話の不要な繰り返し。

例) 海に行っていないという文脈において：「S: 海に行かないなあ」

- 質の公準違反（矛盾）

これまでの文脈と矛盾する内容の発話。また、これまでの話題に対する態度（Positive/Negative の極性）が急に転換する発話。

例) 人間関係が薄いという文脈において：「S: 人間関係は深いです」

- 関係の公準違反（無関係話題）

これまでの文脈における話題から逸脱する発話。話の主たる話題ではなく、直前のユーザ発話中の話題語でない単語から話を展開させてしまうような場合が典型例。

例) 定期券を無くすという文脈において：「U: 首からかけておくとか？」→「S: 左右に振りますよねえ」

- 様態の公準違反（関連性不明）

話題は継続しており、直前発話との関係だけで見ると意味が通るように見えるが、それ以前の文脈を踏まえると、その発話の意図が理解できない状況。

例) 薬の話題において：「S: 薬は大丈夫ですね」

- 話題展開への不追随（不追随）

ユーザからの話題の展開が起こっているにも関わらず、それに追随できていない状況。

例) ドイツの話題の後：「U: あなたは海外経験ありますか？」→「S: ドイツは小売価格が高いです」

### 5.1.4 「環境」の小分類

会話の文脈以外の原因で生じた破綻を分類する。

- 無根拠（共通基盤欠如）

根拠のない、一方的な主張。

例) 「S: マグロは鮮度が悪いですよええ」

- 矛盾（一般常識欠如）

一般的に正しいと信じられている常識と矛盾する発話。

例) 「S: 熱中症はいいですね」

- 非常識（社会性欠如）

罵詈雑言など、対話相手との社会的関係を破壊する発話。あるいは口調（人格や社会

的属性)が突然変化する発話.

例)「S: プールはいいですね」→「U: 探しとくね」→「S: 知らんのかい」

## 5.2 分類体系草案の課題

考案した分類体系は一見よく纏まっており、それなりの一致度で分類を行えることが期待できた。そこで、破綻アノテータが付けたコメントを参考にしながら、タスク参加者で予備的に破綻の分類を行ってみた。しかしながら、予想以上にアノテータ間で一致しないことがわかった( $\kappa$  値で 0.1 から 0.3 程度の範囲)。個々人の主観に任せた破綻アノテーションでは低めの一致度でもよいが、破綻の分類についてはなるべく客観性の高い分類ができることが望ましい。

破綻の分類においてアノテータ間の不一致が大きい原因が、主にアノテーションの手順や教示、アノテータの訓練不足などにあるのか、それとも分類体系自体にあるのか、まだはっきりしていないが、少なくとも以下のような課題が分かっている。

- 検討に際しての分類作業は排他的に一発話・一分類で行ったが、複数の大分類に渡ると思われる破綻がいくつか見られた。例えば、非文・発話対制約違反・話題からの逸脱のように、複数の大分類に渡る破綻が同時に起こることがあり得る。
- 発話の意味制約違反については、典型的な例は「発話」レベルのものと判断しやすいが、解釈次第であることも多い。例えば、「仕事は真面目ですね」という発話は、「仕事」を一般的な概念として捉えれば意味制約違反と判断できるが、ある個人の「業績・仕事ぶり」を意味すると解釈すれば、発話のレベルでは問題がないことになる。「文脈という概念を持ち込むと、文の意味と発話(話し手)の意味を区別することはもはやできない(Levinson 2000)」という見方に立てば、そもそも意味制約違反の小分類を「発話」のレベルに設けることが不適切かもしれない。
- 誤解は、直前の発話に対するものという定義から「応答」の大分類に含めていたが、実際には文脈まで見ないと誤解とは言えない場合も見つかった。これも「応答」でなく「文脈」に含めるか、あるいは「応答」「文脈」の両方に設ける必要があると思われる。
- 分類の問題というよりは、多分に破綻の認定自体の問題であるが、読み手側の知識不足や、表現に対する不慣れによって解釈できなかったため、破綻とされていることもある。例えば、「みんっ」という発話は、意味のある表現に解釈できない人と、「見ない」という意味に解釈できる人がいる。この場合、結果的に、破綻の分類も人により異なってくる。
- 「応答」「文脈」のレベルに導入した Grice の公準に基づく分類は、特に一致率が低かった。これは現状のシステムが出力する発話が、自分のことなのに伝聞で話すなどの不自然な様態や、対話相手のキャラクタが突然変わるなど、通常の人同士の対話で見られないようなものであるために、解釈が難しいことも一因であると考えている。Grice の公準に基づく類型化は、典型例の整理・説明には有用であっても、あまり典型的ではない破

綻の分類には適していない可能性がある。そうだとすれば、小分類のレベルで、各公準違反を事例別にさらに細分化するか、あるいは別の視点での分類を用意する必要がある。

## 6 関連研究

本研究では、非課題指向型対話（雑談対話）に焦点を絞っているが、課題指向型対話システムの文脈では対話システムのエラー分析は活発に行われてきており、いくつかのエラーの分類体系が提案されている。

まず、Clark の提案するコミュニケーション階層モデルに基づくエラーの分類体系 (Clark 1996) が挙げられる。Clark によれば、コミュニケーションのエラーは4つのレベルからなっている。チャンネルレベル、信号レベル、意図レベル、会話レベルである。チャンネルレベルとはやり取りが開始されているかどうかに関わる。信号レベルとはシンボルのやり取りに関わり、意図レベルは対話相手の意図の認識に関わる。会話レベルは、共同行為に関わるものである。下位レベルのエラーが起きていれば、上位レベルでもエラーとなり (upward causality), 上位レベルにエラーがなければ、下位レベルにエラーがないとされる (downward evidence)。このような階層に基づいて、会議室予約システムの不理解によるエラーを分析するという研究がなされている (Bohus and Rudnicky 2005)。また、スマートホームとレストラン情報案内というドメインにおいて、同様の分析もなされている (Möller, Engelbrecht, and Oulasvirta 2007)。Paek は Clark の4つの階層が対話システムのエラー分析に一般性を持っているということを、教育や医療といった複数分野での対話分析の事例から議論している (Paek 2003)。

本論文では Grice の公準 (Grice 1975) をエラーの類型化に用いているが、課題指向型対話システムのエラー分析においても Grice の公準は利用されてきた。Dybkjær et al. (Dybkjær et al. 1996) および Bernsen et al. (Bernsen et al. 1996) はフライト情報案内システムのエラー分析を Grice の公準および独自の対話分析から得られた知見をもとにエラーの類型化を行っている。たとえば、Grice の公準以外の要素として、対話の非対称性、背景知識、メタ対話能力に関わるエラーが挙げられている。電話応答システムにおける対話評価の観点として、Grice の公準に基づく要素を導入することも提案されている (Möller 2005)。

特定のモデルや理論をベースにするのではなく、特定のシステムや対話ドメインの対話を綿密に分析することによりエラーを類型化した例も多い。Aberdeen and Ferro はフライト情報案内システムの分析により、命令に応答しない、何度も同じプロンプトを表示するなどのエラーに類型化している (Aberdeen and Ferro 2003)。また、Green らによって対話機能を持つサービスロボットについてもエラー分析がされており、ロボットに特有のエラーとして、動作と発話のタイミングがずれるというエラーや、指さしなどのポインティング動作のエラーなどが独自のカテゴリとして分類されている (Green, Eklundh, Wrede, and Li 2006)。Dzikovska らは、教育

対話システム (tutoring system) のエラーの類型化を行っている (Dzikovska, Callaway, Farrow, Moore, Steinhäuser, and Campbell 2009).

対話システムはいくつかのモジュールから構成される. このため, エラーの類型化の一つの方法として, エラーを起こしたモジュールがどれかによって分類する研究もある (Ward, Rivera, Ward, and Novick 2005). たとえば, 音声認識, 音声理解, 発話生成, 音声合成といった単位でエラーを類型化する. 音声認識によるエラーが多ければ, 音声認識モジュールを改善すればよいという方針に繋がる. モジュール構成が明確で, 各モジュールのエラーが比較的独立と考えられるのであれば, このような類型化の手法は有効である.

本研究の類型化の手順は (Dybkjær et al. 1996) のものに近い. Grice の公準を用いながら, 対話コーパスについて独自の分析を行いエラーを類型化しているからである. 本研究と Dybkjær らのものとの違いは, 本研究が雑談対話システムを扱っていることである. 課題指向型対話システムに比べタスクやドメインの制約が少ない雑談対話において, エラーの定義, どのようなエラーが起こりうるかは把握されてこなかった. 本研究は, そのような背景に基づき, 雑談対話コーパスの作成およびその類型化を行ったものである. なお, 本研究では Clark の階層モデルは用いていない. これは, 主にテキスト対話を扱っていることによる. テキストのやり取りであれば, チャネルレベルと信号レベルのやり取りは基本的に担保されており, 残りの二つの階層のみに基づいて分類をすることになる. 雑談対話の内容の複雑さを鑑みればこの粒度は粗い. また, モジュールごとにエラーを分析する方法論についてであるが, 雑談対話システムの構成は複雑であり, 単体のモジュールにエラーの分析を起因させることは難しい. また, エラー分析として対話システムの内部構造に立ち入らない方が, システムに依らないエラー分析が可能であり, 特定のシステムに依存しない, 汎用性の高いエラーの類型化が期待できる.

なお, 雑談対話システムのエラー分析は, 対話破綻の自動検出につながるものとして期待されている. 自動検出ができれば, 対話システムが自身の発話を行う前に, その発話に問題があれば, 別の発話候補に切り替えるといったことが可能になる. また, 何らかのエラーを伴う発話をしてしまった後に, 自身の誤りに気づいて, それを訂正するといったことも可能となる.

課題指向型の音声対話システムの文脈では, 音声認識, 発話理解, 対話管理などの各モジュールから得られる特徴量から対話に破綻が起きているかどうかを判定する手法がいくつか提案されている. たとえば, Walker ら (Walker, Langkilde, Wright, Gorin, and Litman 2000) や Herm ら (Herm, Schmitt, and Liscombe 2008) は, コールセンタにおける通話について, 問題が起こっているかどうかを数ターンで判定する判定器を機械学習の手法で構築している. 対話中のユーザの満足度の遷移を推定する研究もされている (Schmitt, Schatz, and Minker 2011). これらは雑談対話を扱っていないが, 目的意識は本論文での取り組みと近い.

雑談対話においては, Chai らがユーザの対話行為の系列の情報を用いて, 問題のある質問応答ペアかどうかの判別を行っている (Chai, Zhang, and Baldwin 2006). Xiang らは, 対話行為に加

え, 感情の系列を用いることで, 雑談対話における問題発話の検出を行っている (Xiang, Zhang, Zhou, Wang, and Qin 2014). Higashinaka らも, 雑談対話システムの発話の結束性をさまざまな素性から推定する手法を提案している (Higashinaka, Meguro, Imamura, Sugiyama, Makino, and Matsuo 2014b). しかしながら, これらの研究は精度がいまだ高いとは言えず, また, 対話破綻の類型化なども行われていない. 今後エラー分析を詳細に行うことで, 対話破綻の原因を明らかにし, 高精度な破綻検出を実現したいと考えている.

## 7 おわりに

本論文では, 雑談対話におけるエラー分析にむけた人・機械間の雑談対話コーパスの構築と対話破綻のアノテーションについて報告した. そして, 構築したコーパスに含まれる破綻を分析し, 考案した破綻の分類体系について議論した.

アノテーション方法の開発にあたっては, 破綻の認定における主観性の高さを認めつつ, 許容可能な範囲のコストで, 客観的な分析の対象となりうる有用なデータを得られるように, 著者らを含む 15 拠点からの研究者で議論・試行し, 工夫を施した. 今回報告した方法と結果は, 破綻に関する今後のコーパス構築に限らず, 同じように主観性の高い別種の言語現象についてのコーパス構築においても手法開発の参考として寄与するものと考えられる.

構築したコーパスでは, 対話を破綻させているシステムの発話に対して, 複数の作業者によってラベルとコメントが付与されている. 破綻の判断については, 事細かなガイドライン・判定方法は示さず, 各個の主観に基づいたアノテーションを行った. このためアノテーションの一致率はそれほど高くないが, システムとの対話に対して人間が不満を持つ点, 持たない点, その個人差について, 興味深いデータを収集できた. また, アノテータ間の一致についての分析からは, 破綻でない発話よりも破綻発話のほうが判定が揺らぎがちであること, アノテータが大きな傾向の違いを持つグループに分かれる可能性があること, などが明らかになった.

一方で, 今回のコーパス構築手法には, 改善の余地があることも確かである. 破綻の判定が揺らぐ要因の 1 つとして, ユーザが想定した対話相手のイメージの違いが存在する. 今回は「待合室や飛行機などで隣り合った見知らぬ人」とだけ指定したが, 性別・年齢・性格など, ユーザができる想定には依然大きな自由度があった. 例えば, 子供染みた発言や冗談は, 想定する相手によって許容できる範囲が変わってくる. 今後のデータ収集においては, 対話相手のイメージをもっと細かくユーザに指定する, あるいは対話前にユーザが想定した相手のイメージ, 対話後に残った相手のイメージを, 対話データと同時に収集すると, より踏み込んだ分析が可能になるだろう.

アノテーションにおける, 破綻を含む先行文脈の扱いについても, さらなる検討が望まれる. 例えば, 今回は破綻があったところで対話がリセットされたとはせず, 破綻も含めて先行文脈



として作業を行うように指示をした。これにより、会話が進んでいけばいくほど破綻が認定されやすくなった可能性があるが、一度破綻したことで文脈上の制約が減り破綻が認定されにくくなっていた可能性もある。

これまで人・機械の雑談対話を体系的に収集し、整備したコーパスは存在せず、今回の収集は初の試みである。今回構築したコーパス中の雑談対話は、1つの雑談システムだけを用いて収集したものであるので、破綻の種類の網羅性やその分布の普遍性について言えることには限りがあるが、システム構築に使用した雑談 API は (Higashinaka, Imamura, Meguro, Miyazaki, Kobayashi, Sugiyama, Hirano, Makino, and Matsuo 2014a) に基づく、現時点で最も複雑な雑談システムの1つであり、少なくとも網羅性については他のシステムを利用した場合と同等かそれ以上、確保できていると考えている。今後、他の雑談システムを使い、本論文で示した方法でデータの収集とアノテーション・分析を行っていくことで、破綻の分布の普遍性を高め、現在の雑談技術・自然言語処理技術が抱える課題により深くアプローチできると期待している。

本稿で示した破綻の分類体系の草案にはまだ改善しなければならない点があるが、破綻の種類を事例的に整理したことで、雑談対話で起こりうる問題について一定の見通しを示すことができた。雑談対話において破綻の種類を分類しようする際に何が問題となるのかを明らかにしたことも、今回の取り組みで得た成果の1つである。

今回構築したコーパスは、破綻検出技術の開発・評価データとして利用することができる<sup>3</sup>。雑談システム自体はそれぞれの目的や利用状況、対象ユーザの想定などが異なるため、直接に比較することが難しく、システム内部の技術的課題について研究者間で議論することが難しい。しかし、雑談システムの入出力であるテキストだけを対象とし、複数の機関が並行して共通のデータで破綻検出技術について開発とエラー分析を進めれば、より一般性の高い議論ができるし、そこから各々の雑談システム自体の技術課題に対しても知見を得られるだろう。また開発された破綻検出技術は、それ自体、多くの研究者・開発者にとって有用なツールを提供できるだろう。

今後は、別システムでのデータの収集や破綻の分類体系の改良を行いながら、破綻検出技術の研究を進めていきたい。

<sup>3</sup> 本論文掲載時点で、コーパスは次の URL で公開されている：

<https://sites.google.com/site/dialoguebreakdown-detection/chat-dialogue-corpus>

また、本コーパスの公開にあわせて開催された破綻検出チャレンジの結果が、(東中, 船越, 小林, 稲葉 2015) にまとめられている。

## 謝 辞

対話データの収集, および, 対話破綻アノテーションにご協力頂いた Project Next NLP 対話タスクの拠点参加者とその関係者の皆さま, 対話データ収集のためのシステム構築とサーバ運営にご協力いただいた広島市立大の稲葉通将氏に感謝いたします. システム構築には株式会社 NTT ドコモの雑談対話 API を使わせていただきました.

本稿の著者は, タスク共同リーダー2名と, 5節の類型化に直接的に貢献したワーキンググループのメンバに限っていますが, その他の拠点参加者の方々におかれても, 電話会議やメーリングリストでの議論を通じて本稿の執筆に様々に貢献していただきました. 一人一人お名前を挙げるのは控えさせていただきますが, 改めて拠点参加者の皆さまのご協力にお礼申し上げます.

最後になりますが, 有益なコメントをいただいた編集委員・査読者の皆さまにお礼申し上げます.

## 参考文献

- Aberdeen, J. and Ferro, L. (2003). “Dialogue Patterns and Misunderstandings.” In *Proceedings of ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pp. 17–21.
- Banchs, R. E. and Li, H. (2012). “IRIS: A Chat-oriented Dialogue System Based on the Vector Space Model.” In *Proceedings of the ACL 2012 System Demonstrations*, pp. 37–42.
- Bang, J., Noh, H., Kim, Y., and Lee, G. G. (2015). “Example-based Chat-oriented Dialogue System with Personalized Long-term Memory.” In *Proceedings of BigComp*, pp. 238–243.
- Bernsen, N. O., Dybkjær, H., and Dybkjær, L. (1996). “Principles for the design of cooperative spoken human-machine dialogue.” In *Proceedings of ICSLP*, Vol. 2, pp. 729–732.
- Bickmore, T. W. and Cassell, J. (2001). “Relational Agents: A Model and Implementation of Building User Trust.” In *Proceedings of CHI*, pp. 396–403.
- Bohus, D. and Rudnicky, A. I. (2005). “Sorry, I Didn’t Catch That!—An Investigation of Non-understanding Errors and Recovery Strategies.” In *Proceedings of SIGDIAL*, pp. 128–143.
- Chai, J. Y., Zhang, C., and Baldwin, T. (2006). “Towards Conversational QA: Automatic Identification of Problematic Situations and User Intent.” In *Proceedings of COLING/ACL*, pp. 57–64.
- Clark, H. H. (1996). *Using Language*. Cambridge University Press.
- Dybkjær, L., Bernsen, N. O., and Dybkjær, H. (1996). “Grice Incorporated: Cooperativity in Spoken Dialogue.” In *Proceedings of COLING*, Vol. 1, pp. 328–333.
- Dzikovska, M. O., Callaway, C. B., Farrow, E., Moore, J. D., Steinhauser, N., and Campbell,

- G. (2009). “Dealing with Interpretation Errors in Tutorial Dialogue.” In *Proceedings of SIGDIAL*, pp. 38–45.
- Green, A., Eklundh, K. S., Wrede, B., and Li, S. (2006). “Integrating Miscommunication Analysis in Natural Language Interface Design for a Service Robot.” In *Proceedings of IEEE/RSJ*, pp. 4678–4683.
- Grice, H. P. (1975). “Logic and Conversation.” In Cole, P. and Morgan, J. (Eds.), *Syntax and Semantics 3: Speech Acts*, pp. 41–58. New York: Academic Press.
- Herm, O., Schmitt, A., and Liscombe, J. (2008). “When Calls Go Wrong: How to Detect Problematic Calls Based on Log-files and Emotions?” In *Proceedings of INTERSPEECH*, pp. 463–466.
- 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将 (2015). 対話破綻検出チャレンジ. 言語・音声理解と対話処理研究会 第 75 回研究会 (第 6 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-75-B502 巻, pp. 27–32.
- Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014a). “Towards an Open-domain Conversational System Fully Based on Natural Language Processing.” In *Proceedings of COLING*, pp. 928–939.
- Higashinaka, R., Meguro, T., Imamura, K., Sugiyama, H., Makino, T., and Matsuo, Y. (2014b). “Evaluating Coherence in Open Domain Conversational Systems.” In *Proceedings of INTERSPEECH*, pp. 130–133.
- 稲葉通将, 神園彩香, 高橋健一 (2014). Twitter を用いた非タスク指向型対話システムのための発話候補文獲得. 人工知能学会論文誌, **29** (1), pp. 21–31.
- 小林優佳, 山本大介, 大内一成, 長健太, 瀬戸口久雄, 土井美和子 (2011). 安心でワクワクさせるロボット対話インタフェースを目指して: 対話とセンサによる高齢者の健康情報収集. 電子情報通信学会技術研究報告 クラウドネットワークロボット, 111 巻, pp. 11–16.
- Landis, J. R. and Koch, G. G. (1977). “The Measurement of Observer Agreement for Categorical Data.” *Biometrics*, **33**, pp. 159–174.
- Levinson, S. C. (2000). *Presumptive Meaning: The Theory of Generalized Conversational Implicature*. MIT. (邦訳: 「意味の推定: 新グライス学派の語用論」, 研究社, 2007) .
- Matsuyama, Y., Akiba, I., Saito, A., and Kobayashi, T. (2013). “A Four-Participant Group Facilitation Framework for Conversational Robots.” In *Proceedings of SIGDIAL*, pp. 284–293.
- Möller, S. (2005). “Parameters for Quantifying the Interaction with Spoken Dialogue Telephone Services.” In *Proceedings of SIGDIAL*, pp. 166–177.
- Möller, S., Engelbrecht, K.-P., and Oulasvirta, A. (2007). “Analysis of Communication Failures for Spoken Dialogue Systems.” In *Proceedings of INTERSPEECH*, pp. 134–137.

- 大西可奈子, 吉村健 (2014). コンピュータとの自然な会話を実現する雑談対話技術. NTT DoCoMo テクニカル・ジャーナル, **21** (4), pp. 17–21.
- Paek, T. (2003). “Toward a Taxonomy of Communication Errors.” In *Proceedings of ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pp. 53–58.
- Schmitt, A., Schatz, B., and Minker, W. (2011). “Modeling and Predicting Quality in Spoken Human-computer Interaction.” In *Proceedings of SIGDIAL*, pp. 173–184.
- 関根聡 (2015). Project Next NLP 概要 (2014/3-2015/2). 言語処理学会第 21 回年次大会ワークショップ：自然言語処理におけるエラー分析.
- Walker, M., Langkilde, I., Wright, J., Gorin, A., and Litman, D. (2000). “Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You?” In *Proceedings of NAACL*, pp. 210–217.
- Wallace, R. S. (2004). “The Anatomy of A.L.I.C.E.” Tech. rep., A.L.I.C.E Artificial Intelligence Foundation, Inc.
- Ward, N. G., Rivera, A. G., Ward, K., and Novick, D. G. (2005). “Root Causes of Lost Time and User Stress in a Simple Dialog System.” In *Proceedings of INTERSPEECH*, pp. 1565–1568.
- Wilcock, G. and Jokinen, K. (2013). “Wikitalik Human-robot Interactions.” In *Proceedings of ICMI*, pp. 73–74.
- Xiang, Y., Zhang, Y., Zhou, X., Wang, X., and Qin, Y. (2014). “Problematic Situation Analysis and Automatic Recognition for Chinese Online Conversational System.” In *Proceedings of CLP*, pp. 43–51.

## 略歴

東中竜一郎：1999 年慶應義塾大学環境情報学部卒業, 2001 年同大学大学院政策・メディア研究科修士課程, 2008 年博士課程修了. 2001 年日本電信電話株式会社入社. 現在, NTT メディアインテリジェンス研究所に所属. 質問応答システム・音声対話システムの研究開発に従事. 博士 (学術). 言語処理学会, 人工知能学会, 情報処理学会, 電子情報通信学会各会員.

船越孝太郎：2000 年東京工業大学工学部情報工学科卒業. 2002 年同大学大学院情報理工学研究科計算工学専攻修士課程修了. 2005 年同博士課程修了. 同年同大学院特別研究員. 2006 年より株式会社ホンダ・リサーチ・インスティテュート・ジャパン入社. 2013 年より同シニア・リサーチャ. 博士 (工学). 自然言語理解, マルチモーダル対話に関する研究に従事. 情報処理学会, 人工知能学会, 言語処理学会, ヒューマンインタフェース学会, ACM SIGCHI 各会員.

**荒木 雅弘**：1988 年京都大学工学部卒業。1993 年京都大学大学院工学研究科博士課程研究指導認定退学。京都大学工学部助手、同総合情報メディアセンター講師を経て、現在京都工芸繊維大学大学院工芸科学研究科准教授。音声対話システムおよびマルチモーダル対話記述言語の研究に従事。ACL、ISCA、情報処理学会等各会員。博士（工学）。

**小林 優佳**：2004 年東京工業大学大学院理工学研究科修士課程終了（機械制御システム専攻）。同年東芝家電製造株式会社（現東芝ライフスタイル株式会社）入社。2008 年株式会社東芝研究開発センター入社。音声対話システムの研究開発に従事。電子情報通信学会、情報処理学会、人工知能学会各会員。

**塚原 裕史**：1994 年中央大学理工学部物理学科卒業。1996 年同大学大学院博士課程前期修了。1999 年同大学院博士課程後期修了。博士（理学）。2000 年日立ソフトウェアエンジニアリング株式会社入社。分散オブジェクト地理情報システムの研究・開発に従事。2005 年株式会社デンソーアイティラボラトリ入社。現在同社研究開発グループ勤務。自動車向け人工知能応用システムに関する研究・開発に従事。日本物理学会、情報処理学会各会員。

**水上 雅博**：2012 年同志社大学理工学部卒業。2014 年奈良先端科学技術大学院大学情報科学研究科修士課程修了。同年より同大学院博士後期課程在学。自然言語処理および音声対話システムに関する研究に従事。人工知能学会、音響学会、言語処理学会各会員。

（2015 年 5 月 21 日 受付）

（2015 年 8 月 6 日 再受付）

（2015 年 11 月 3 日 採録）